STATISTICS IN MEDICINE
Statist. Med. 2003; **22**:3369–3381 (DOI: 10.1002/sim.1638)

Quantifying epidemiologic risk factors using non-parametric regression: model selection remains the greatest challenge[‡]

Philip S. Rosenberg[1,*,†], Hormuzd Katki[1], Christine A. Swanson[2], Linda M. Brown[1], Sholom Wacholder[1] and Robert N. Hoover[2]

[1] *National Cancer Institute, Division of Cancer Epidemiology and Genetics, Biostatistics Branch, 6120 Executive Boulevard, EPS 7006, Rockville, MD 20852, U.S.A.*
[2] *National Cancer Institute, Epidemiology and Biostatistics Program, 6120 Executive Boulevard, EPS 8094, Rockville, MD 20852, U.S.A.*

SUMMARY

Logistic regression is widely used to estimate relative risks (odds ratios) from case–control studies, but when the study exposure is continuous, standard parametric models may not accurately characterize the exposure–response curve. Semi-parametric generalized linear models provide a useful extension. In these models, the exposure of interest is modelled flexibly using a regression spline or a smoothing spline, while other variables are modelled using conventional methods. When coupled with a model-selection procedure based on minimizing a cross-validation score, this approach provides a non-parametric, objective, and reproducible method to characterize the exposure–response curve by one or several models with a favourable bias–variance trade-off. We applied this approach to case–control data to estimate the dose–response relationship between alcohol consumption and risk of oral cancer among African Americans. We did not find a uniquely 'best' model, but results using linear, cubic, and smoothing splines were consistent: there does not appear to be a risk-free threshold for alcohol consumption *vis-à-vis* the development of oral cancer. This finding was not apparent using a standard step-function model. In our analysis, the cross-validation curve had a global minimum and also a local minimum. In general, the phenomenon of multiple local minima makes it more difficult to interpret the results, and may present a computational roadblock to non-parametric generalized additive models of multiple continuous exposures. Nonetheless, the semi-parametric approach appears to be a practical advance. Published in 2003 by John Wiley & Sons, Ltd.

KEY WORDS: epidemiologic methods; logistic models; oral cancer; alcohol drinking

INTRODUCTION

In 1973, Seigel and Greenhouse published a seminal paper demonstrating the validity and usefulness of analysing data from case–control studies using the logistic regression model [1].

[*] Correspondence to: Philips S. Rosenberg, National Cancer Institute, Division of Cancer Epidemiology and Genetics, 6120 Executive Boulevard, EPS 7006, Rockville, MD 20852, U.S.A.

[†] E-mail: philip_rosenberg@nih.gov

[‡] This article is a U.S. Government work and is in the public domain in the U.S.A.

Received May 2003
Accepted May 2003

Previously, logistic regression had been used to analyse data from cohort studies and clinical trials, but the sole method available to estimate relative risks (odds ratios) from case–control studies dichotomized each risk factor to construct 2×2 tables relating the dichotomized risk factor to disease.

Following publication of this work, logistic regression has become a ubiquitous method for analysing case–control data, and it has been extensively refined. In this paper, we describe an extension of the standard logistic regression model that allows one to analyse the effects of a continuous exposure non-parametrically. In theory, these refined analyses are less sensitive to misspecification of the exposure–response relationship, an issue of concern to epidemiologists [2,3]. In a sense, this approach provides the ultimate level of refinement, because a relative risk is estimated in a locally adaptive manner for each infinitesimal increment in exposure. However, such exquisite refinement begs the question: Are these advanced methods 'gilding the lily' [4]? Standard methods allow one to estimate relative risks within categories of exposure or, alternatively, to fit smooth parametric curves. Aren't these models sufficient?

Our interest in non-parametric logistic regression was spurred by substantive epidemiologic studies conducted by investigators in the National Cancer Institute's Division of Cancer Epidemiology and Genetics. They had completed a series of population-based case–control studies to help determine why African Americans have a higher risk of certain cancers, including oral cancers, than Americans of European descent [5]. Although basic patterns of risk could be discerned using standard methods, the numbers of African American cases were comparatively small (because the population is smaller), and dose–response relationships were correspondingly less well characterized when the analysis was restricted to them. Furthermore, although oral cancer is widely perceived to be a disease of alcohol and tobacco abuse, the relative risks associated with low and moderate levels of exposure remain unclear.

In this paper, we review some modern methods of non-parametric risk regression, and apply these methods to case–control data to estimate the dose–response relationship between alcohol consumption and risk of oral cancer in African Americans. Through this example, we hope to demonstrate that these refined analytical methods are worthy of broader application and continued development, not just because of their statistical elegance, but also because they can help answer key scientific questions [6]. However, as our example illustrates, identification of the best-fitting model is not automatic, and results from the modelling procedure require careful interpretation.

MODELS AND DATA

The oral cancer study

The National Cancer Institute's Oral Cancer Study was a large population-based case–control study [5]. It enrolled 1065 incident cases of primary oral cancer diagnosed between January 1984 and March 1985 in Los Angeles, Santa Clara and San Mateo California, metropolitan Atlanta, and the state of New Jersey. Controls were frequency matched to cases by age, sex and race. The analysis presented here is restricted to the African American arm of the study (194 cases and 203 controls). The primary exposure measure is the usual number of 1 oz ethanol-equivalent drinks consumed per week as inferred from questionnaire data; this measure treats 12 oz of beer, 4 oz of wine and 1.5 oz of liquor as equivalent units of exposure. The estimates we present are adjusted for sex, quartiles of age (21–48, 49–56, 57–65 and 66–80 years of age) and recent cigarette smoking (non-smoker, 1–19, 20–39, and 40+ cigarettes per day).

Additive models

Let the 0–1 response variable (case–control status) be $Y_i \sim \text{Bernoulli}(\theta_i)$, $i=1,\ldots,n$, and for each observation let $x_i'=(x_{1i},\ldots,x_{pi})$ be an associated row vector of covariates. A generalized additive model for θ_i is

$$\text{logit}\,\theta_i=\alpha+\sum_{j=1}^{p} f_j(x_{ji})$$

where $f_j(.)$ are arbitrary smooth functions to be estimated [7]. In principle, the specific form of the $f_j(.)$ can be estimated non-parametrically using a cross-validation function. However, as we will illustrate in Result section, it may be difficult in general to estimate all of the $f_j(.)$ non-parametrically because the cross-validation function may have multiple local minima [8]. Therefore, a reasonable alternative is to fit a restricted model, called a semi-parametric generalized linear model [9], in which one of the continuous variables, called the splined variable, is treated non-parametrically, while the other $p-1$ variables are modelled using standard parametric approaches. In our analysis, the number of drinks per week will be the splined variable, while sex, age group and cigarette smoking will be treated using standard approaches. Let the splined variable be the pth variable, denoted by t, and let $x_{0i}'=(x_{1i},\ldots,x_{p-1,i})$ be the remainder. A semi-parametric generalized linear model for θ_i is

$$\text{logit}\,\theta_i=x_{0i}'\beta+f(t_i)$$

where β is a vector of regression coefficients and $f(t)$ is an arbitrary smooth function to be estimated. The intercept α is absorbed into $f(t)$. Although standard methods can be very flexible when the x_{0i} encode flexible functions of the data, e.g. include linear and quadratic terms, etc., we treat the splined variable t in a special way to estimate its effects non-parametrically.

Spline functions are piecewise polynomials. A spline is defined by the order of the polynomial pieces, the set of knots or join-points that define a contiguous set of intervals covering the domain of the function and the specific continuity constraints imposed at each knot. A rich algebra permits one to construct splines with desirable properties; the classic work of de Boor presents a general treatment [10]. Hastie and Tibshirani [7] and Green and Silverman [9] provide lucid developments of spline functions from a statistical perspective.

There are two general approaches to model the effects of the splined variable t. One approach models $f(t)$ as a regression spline, the other as a smoothing spline. Each class of spline functions has technical strengths and limitations, and we consider them to be complementary.

Regression splines

A large class of regression splines can be represented using B-spline basis functions, and this basis has computational advantages [10]. However, we can represent the most commonly used types of splines using a simpler algebra based on truncated polynomials [11]. Let $[t_0,t_k]$ cover the range of t, let $\{t_0,t_1,\ldots,t_{k-1},t_k\}$ be a non-decreasing sequence of knots, and let $(x)_+=\max(x,0)$ be the 'truncation' function. Typically, the number of internal knots of a regression spline is considerably smaller than the number of distinct t-values in the data, and when $k=1$ the spline reduces to a simple polynomial. The function

$$f_{\mathrm{L}}(t)=\alpha+\beta_0\,t+\beta_1(t-t_1)_{+}+\cdots+\beta_{k-1}(t-t_{k-1})_{+}$$

is a continuous piecewise-linear spline for the interval $[t_0,t_k]$. The function

$$f_{\mathrm{C}}(t)=\alpha+\beta_0t+\beta_1t^2+\beta_2t^3+\beta_{2+1}(t-t_1)_{+}^{3}+\ldots+\beta_{2+k-1}(t-t_{k-1})_{+}^{3}$$

is a continuous piecewise cubic spline for the same interval. The derivative of $f_{\mathrm{L}}(t)$ is piecewise constant with jumps at $\{t_1,\ldots,t_{k-1}\}$, while $f_{\mathrm{C}}(t)$ is a smooth function—both $f_{\mathrm{C}}(t)$ and its first two derivatives are continuous on $[t_0,t_k]$. (The third derivative of $f_{\mathrm{C}}(t)$ is piecewise constant with jumps at $\{t_1,\ldots,t_{k-1}\}$.) Although not usually thought of as splines, step functions are also piecewise polynomials.

Standard methods allow one to estimate the parameters of the regression spline if the appropriate number and location of knots are known *a priori*. This is rarely the case in epidemiological investigations. One strategy for model selection puts an increasing number of knots at successive quantiles of t observed in the control series, e.g. no internal knots, one internal knot at the median, two at the tertiles, three at the quartiles, etc., up to an arbitrary maximum. In practice, the maximum value is anywhere from 6 to 10, but it could be more if the data set supports it. The best-fit model is selected using the Akaike Information Criterion (AIC) [12], a statistic that is closely related to the cross-validation score [13], which provides an estimate of the mean squared error of the procedure [14]. One advantage of this approach is that it is relatively straightforward to implement using standard software packages.

Smoothing splines

Construction of a smoothing spline $f_{\mathrm{S}_\lambda}(t)$ is more technical. Green and Silverman present the details [9]. The parameter $\lambda>0$ is a 'tuning parameter' that will be described below. In brief, one follows the construction used for the cubic regression splines, but enlarges the knot sequence so that it includes every distinct observed value of the splined variable t. Next, one considers the subset of such splines whose second and third derivatives are zero at the end knots t_0 and t_k. This subset defines the so-called natural cubic splines; the end-conditions imply that $f_{\mathrm{S}_\lambda}(t)$ is linear over the two extreme intervals $[t_0,t_1]$ and $[t_{k-1},t_k]$.

For the semi-parametric logistic regression model, $\operatorname{logit}\theta_i=x'_{0i}\beta+f_{\mathrm{S}_\lambda}(t_i)$, and coefficients for the conventional and splined variables can be estimated using a penalized log-likelihood:

$$\ell_p(\beta,f_{\mathrm{S}_\lambda})=\sum_{i=1}^{n}[Y_i[x'_{0i}\beta+f_{\mathrm{S}_\lambda}(t_i)]-\log(1+\mathrm{e}^{x'_{0i}\beta+f_{\mathrm{S}_\lambda}(t_i)})]-\frac{1}{2}\lambda\int_{t_0}^{t_k}f''_{\mathrm{S}_\lambda}(t)^2\,\mathrm{d}t$$

The standard likelihood is augmented by a roughness penalty that measures the integrated squared second derivative of the estimated spline. It is a remarkable fact that the minimizer of this roughness penalty must be a smoothing spline with a knot at each distinct t_i, and this mathematical fact endows the smoothing splines with an explicit optimality criterion. Larger values of λ restrict the flexibility and effective degrees of freedom of the estimated smoothing spline $\hat{f}_{\mathrm{S}_\lambda}(t)$, while smaller values allow the fitted curve more 'wriggle room' to track the data. As $\lambda\to\infty$, $\hat{f}_{\mathrm{S}_\lambda}(t)$ approaches a straight-line fit, while as $\lambda\to 0$, $\hat{f}_{\mathrm{S}_\lambda}(t)$ will attempt to interpolate the adjusted log odds ratio for cases and controls that share each distinct observed value of t (the model may fail to converge for very small values of λ). The integral form of the penalty would appear to introduce computational difficulties, but fortunately one does not have to use numerical integration. Because $\hat{f}_{S_\lambda}(t)$ is a smoothing spline, the penalty can be expressed as a quadratic form involving fitted $\hat{f}_{S_\lambda}$-values at the knots and a kernel that depends only on the knots [9, Section 2.1.2].

In the calculations, it is convenient to let $\hat{f}_{S_\lambda}(t)$ absorb the model's intercept, since the constant term is always in the span of the smoothing spline whatever be the value of λ. This implies that $\hat{f}_{S_\lambda}(t)$ models a so-called 'floating risk'; the log odds ratio contrasting any two values of t is estimated by the corresponding difference of $\hat{f}_{S_\lambda}(t)$-values [15, 16].

An appropriate value of the tuning parameter is seldom known *a priori*, but can be estimated using a cross-validation score to estimate the mean squared error of the procedure. This is presently a limitation of the approach because it is uncertain what is the best way to define the appropriate score. Furthermore, the methods proposed to date while reasonable require extensive programming. We have implemented an approximate generalized cross-validation score based on a linearized form of the deviance

$$\mathrm{GCV}_{\mathrm{ap}}(\lambda) = \sum_{i=1}^{n} w_i \frac{(z_i - (x_{0i}'\hat{\beta} + \hat{f}_{S_\lambda}(t_i)))^2}{\left(1 - \sum_{i=1}^{n} A_{ii}/n\right)^2}$$

where z_i and w_i are the working response variable and iterative weights at the final iteration of the Fisher scoring algorithm for maximizing the penalized likelihood, and A_{ii} are diagonal elements of an $n \times n$ 'hat matrix' that depends on the w_i, x_{0i} and t_i [9, Sections 4.4 and 5.4]. This expression approximates the weighted sum of prediction errors resulting from leaving out each observation in turn and predicting its value from the other $n - 1$ observations.

Model selection

Whichever type of spline is used, the results are most convincing when the model is selected in a disciplined manner. We recommend that the analyst report the best-fit model, and any other models corresponding to local minima of the AIC or cross-validation curve, regardless of whether models based on other knot configurations appear to give more plausible curves than the curve or curves selected by AIC or cross-validation. This does not preclude reporting these other models as well. However, we feel it is important that the analyst reports the set of models that provide a similar bias versus variance trade-off, and also describes the full set of models considered. Otherwise, the procedure will not be reproducible by other investigators with similar data sets, and it may be hard for the field to reach a consensus.

RESULTS

Figure 1 shows nine linear regression splines for the effects of drinks per week on oral cancer risk among African Americans, and Figure 2 shows nine cubic regression splines fitted to the same data. Table I shows AIC values for the 18 models presented in Figures 1 and 2. Figure 3 plots AIC values versus the number of segments for the cubic and linear regression splines. A cubic spline with two segments (one internal knot at the median exposure observed

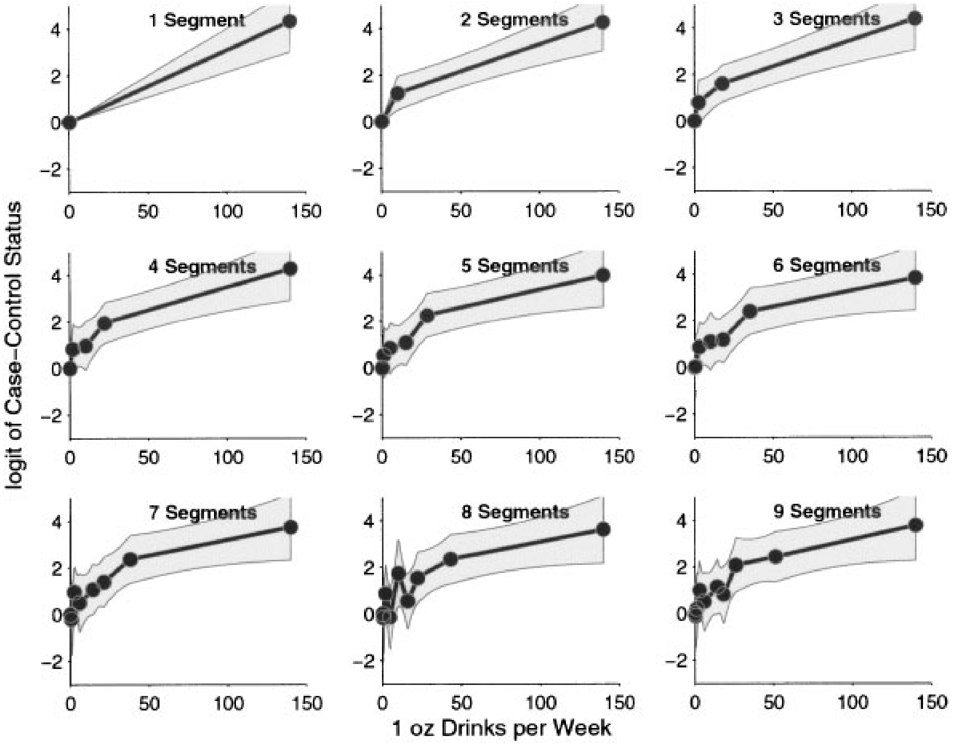

Figure 1. Linear regression spline models of the relationship between alcohol consumption and risk of oral cancer among African Americans, for models with one–nine linear segments. In each panel, the x-axis shows the number of 1 oz drinks per week, and the y-axis shows the log odds ratio for each value of drinks per week relative to non-drinkers. Filled circles show knot locations. Shaded regions show 95 per cent pointwise confidence limits.

in the controls) is the best-fitting cubic regression spline among the set of nine cubic splines considered (Figure 3(a)). However, the AIC curve does not have a convex appearance. The eight-segment models fit substantially better than models with 7 or 9 segments, although not quite as well as the two-segment model. Similar results were obtained for the linear splines (Figure 3(b)): the two-segment linear spline provides the best fit, the eight-segment model fits nearly as well, and it appears to be a local minimum in the sense that it fits substantially better than models with 7 or 9 segments.

Both approaches suggest that the risk of oral cancer increases most rapidly at low exposures. This is apparent by inspection of the fitted log odds ratios, which have the greatest slope at lower compared to higher exposures (Figures 1 and 2). The more flexible cubic spline suggests that at the highest levels of exposure, the rate of increase of the risk slows. There does not appear to be a risk-free threshold for alcohol consumption *vis-à-vis* the development of this tumour: none of the curves has a flat trend at lower exposure levels followed by an increasing trend thereafter.

Figure 4 presents results using the cubic smoothing spline. The GCV score has a global minimum at 14.3 degrees of freedom and a local minimum at 22.76 degrees of freedom (df)

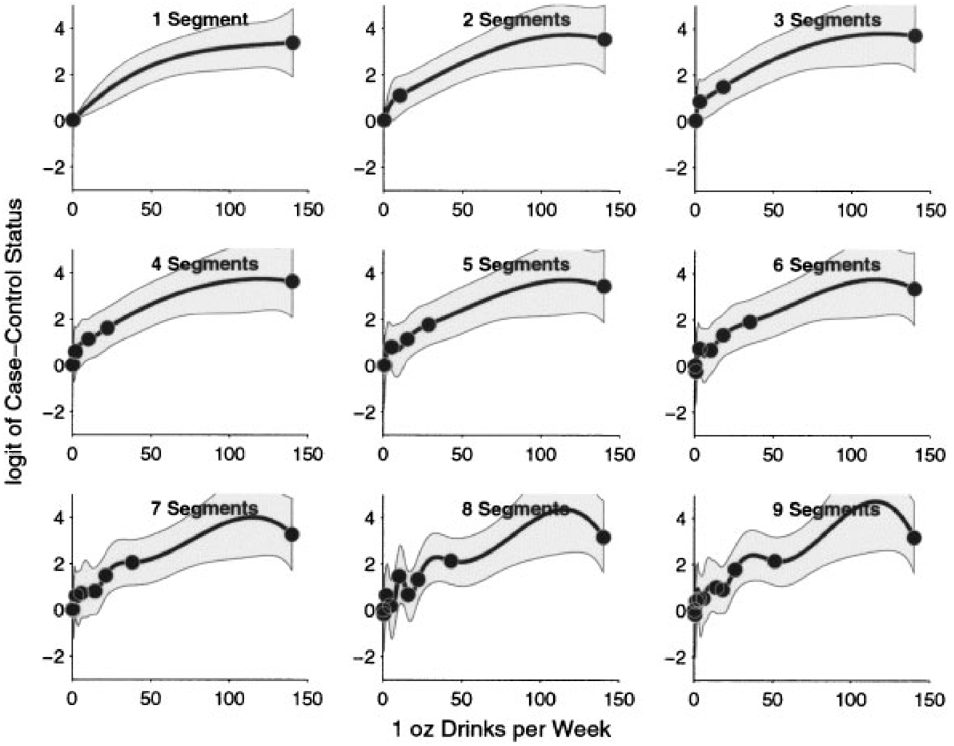

Figure 2. Cubic regression spline models for the relationship between alcohol consumption and risk of oral cancer among African Americans, for models with one–nine cubic segments. See legend to Figure 1 for details.

(Figure 4(a)), including the degrees of freedom used to model sex, age group, and cigarette smoking. Fitted curves for selected values of λ illustrate the increasing flexibility of the fits with increasing λ values (Figure 4(b)). For the best-fit model, the log odds ratio increases linearly up to about 45 drinks per week, and then reaches a plateau until 70 drinks per week (Figure 4(c)); there is a striking increase and decrease in the curve thereafter. This feature is explicable when one examines the distribution of exposure in cases and controls (Figure 4(d)). Less than 5 per cent of controls (nine individuals) consumed more than 70 drinks per week, and as a consequence, the curve above this level is difficult to estimate. The feature in the fitted curve in this region reflects the empirical trend in limited data at the highest exposure levels.

Figure 4(c) also shows 95 per cent pointwise confidence limits with λ fixed at its optimal value based on the GCV curve. Dashed lines show bootstrap confidence limits, while the shaded region shows approximate confidence limits with z_i, w_i and β fixed at values from the final iteration of the Fisher scoring algorithm. The approximate limits derive by smoothing z_i on t_i with weights w_i and applying standard formulae [17]. As expected, the approximate limits are narrower than bootstrap limits because they do not incorporate uncertainty about β. However, they are much easier to compute.

Table I. Akaike information criterion (AIC) values for regression spline models of the relationship between alcohol consumption and risk of oral cancer.

Order	Number of segments	Deviance*	Degrees of freedom (df) for spline	AIC
Linear spline	1	428.03	1	446.03
	2	421.60	2	441.60†
	3	420.29	3	442.29
	4	418.50	4	442.50
	5	417.60	5	443.60
	6	417.05	6	445.05
	7	415.55	7	445.55
	8	410.03	8	442.04
	9	414.67	9	448.67
Cubic spline	1	419.77	3	443.77
	2	418.45	4	442.45†
	3	417.83	5	443.83
	4	418.26	6	446.26
	5	416.48	7	446.48
	6	414.92	8	446.92
	7	413.78	9	447.78
	8	407.74	10	443.74
	9	409.57	11	447.57

* All models include 1 df for the intercept, 1 df for sex, 3 df for age group, and 3 df for cigarette smoking.

† Best-fit model based on AIC criterion.

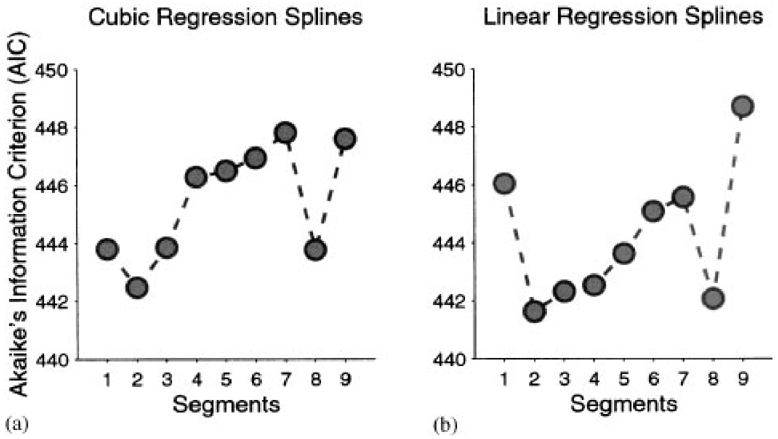

Figure 3. Akaike information criterion (AIC) values for cubic (panel a) and linear (panel b) regression spline models of the relationship between alcohol consumption and risk of oral cancer among African Americans. See Table I for details.

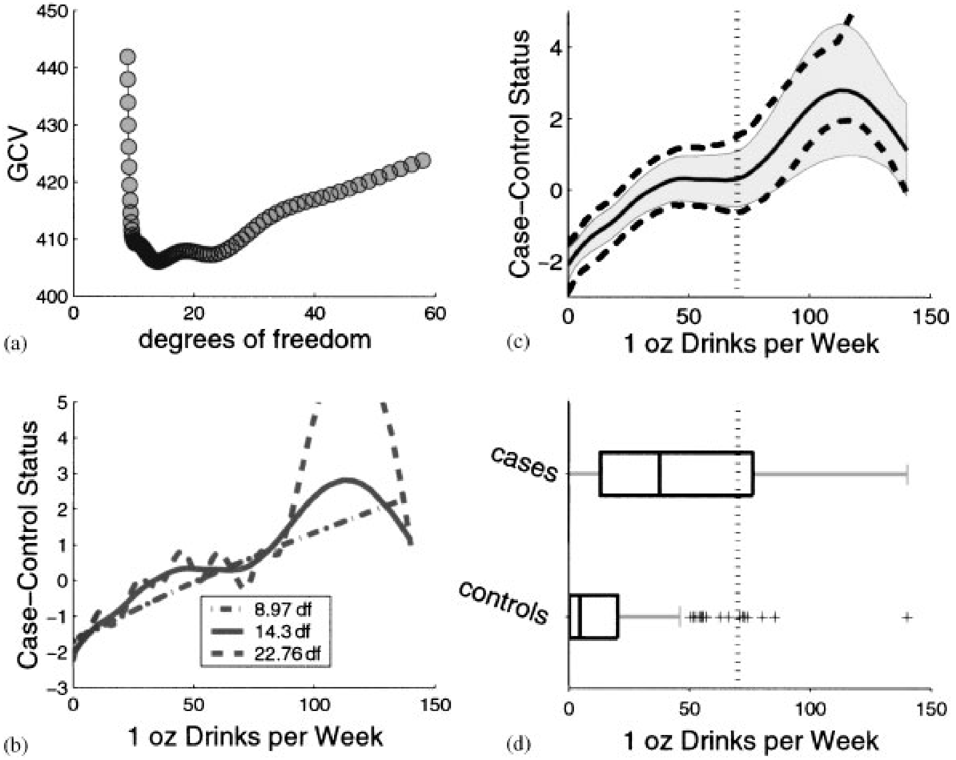

Figure 4. Smoothing spline models of the relationship between alcohol consumption and risk of oral cancer among African Americans. Panel a: generalized cross-validation (GCV) curve. Panel b: Smoothing splines with a low (8.97), optimal (14.3) and high (22.76) number of degrees of freedom. Panel c: Optimal model based on the GCV curve, anticonservative 95 per cent pointwise confidence limits derived using standard formulae (shaded region; see text), and 'complete' limits based on the bootstrap (dashed lines). Panel d: Box plots of the distribution of drinks per week in cases and controls. In panels b–c, *y*-axis values represent 'floating risks:' the log odds ratio contrasting any two values of drinks is found by calculating the difference of the corresponding *y*-values. Spline values above 70 drinks per week are uncertain (see text).

Figure 5 contrasts the best-fitting linear and cubic regression splines and the best-fitting smoothing spline. Including the covariates age, sex and cigarette smoking, these models had 10, 12, and 14.3 df, respectively. The cubic regression spline does not fit significantly better than the linear regression spline (deviances of 418.45 with 378 df for error, versus 421.60 with 380 df for error, respectively), and the AIC values are similar, 442.45 versus 441.60, respectively. The AIC value for the best-fitting smoothing spline (deviance plus twice the non-parametric degrees of freedom) was also similar, 442.29. The models are statistically indistinguishable, and from a substantive perspective, the curves are qualitatively consistent. Each indicates a continuous increase in risk at lower levels of exposure, and none shows evidence for a threshold effect. For comparison, Figure 5 also shows a standard step-function model with the reference category equal to 0 drinks per week and three exposure steps based on tertiles of drinks per week in non-abstaining controls. Compared to the spline fits, the fit

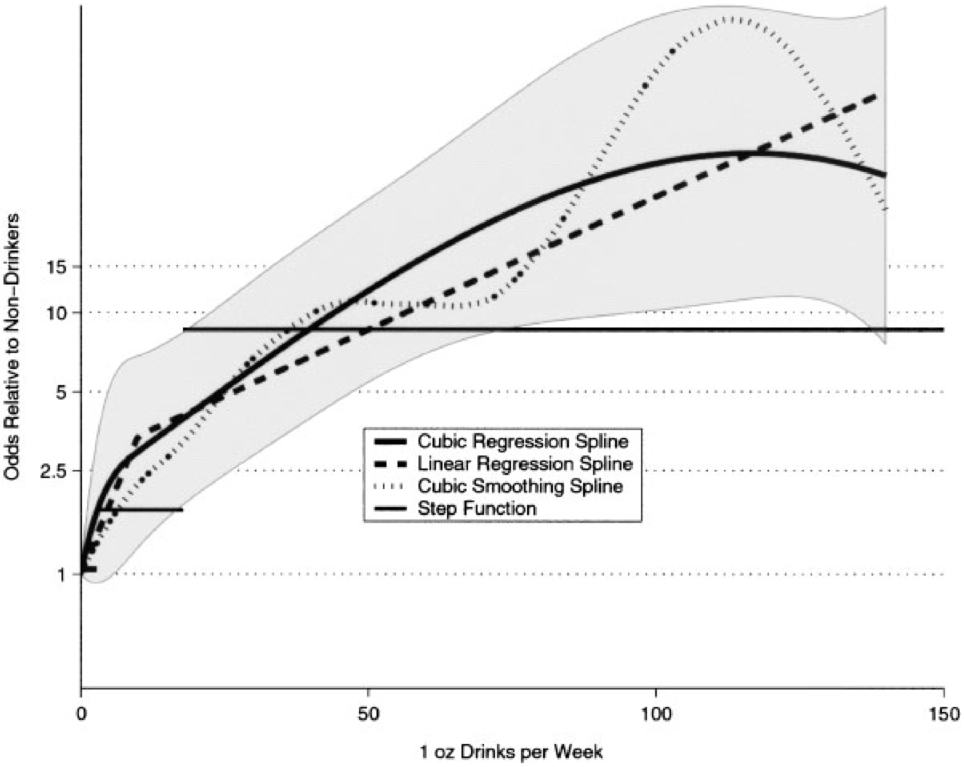

Figure 5. Comparison of estimates of the log odds ratio characterizing alcohol consumption and risk of oral cancer among African Americans. The x-axis shows the number of 1 oz drinks per week, and the y-axis shows the odds ratio for each value of drinks per week relative to non-drinkers on a logarithmic scale. Curves are shown for the best-fitting linear and cubic regression splines selected by AIC, the best-fitting cubic smoothing spline selected by GCV, and a standard step function model with three steps and a reference category of zero. Confidence intervals are shown for the cubic regression spline.

of this step-function model is poor (deviance of 433.90 with 379 df for error, AIC of 455.90), and it does not appear to provide a good approximation to the dose–response curve.

DISCUSSION

Logistic regression models have come of age in the years since Seigel and Greenhouse's seminal paper. One development is the introduction of spline functions that allow one to estimate the risks associated with continuous exposures using very flexible models. These more advanced logistic regression models have been used successfully in diverse applications [16, 18].

In our example, we saw no 'safe threshold' *vis-à-vis* alcohol consumption and the risk of oral cancer among African Americans. This result was not apparent using a standard step-function model. In general, it may be difficult to characterize the dose–response relationship using a standard categorical analysis that divides exposures into ranges: one will not know whether a threshold lies hidden within the span of the first category, the parameter estimates may become unstable as the number of intervals increases and the unavoidable jumps in risk are not biologically plausible. In contrast, in our example, the lack of a threshold in risk at low exposure levels was consistently observed using linear regression splines, cubic regression splines and cubic smoothing splines (Figure 5). Each of these risk models is continuous, and the latter two are smooth. Indeed, perusing Figures 1, 2, and 4, none of the 21 models shown have a strong indication for a threshold. Our analysis was restricted to African Americans, but qualitatively similar patterns were observed among cases and controls of European descent (data not shown).

We have sometimes encountered the sentiment that non-parametric risk regression models are too sensitive for epidemiologic studies because there may be considerable noise in the data relative to the signal. We understand this sentiment but take a more optimistic view. Yes, these methods are very sensitive to the data, but we see this as a plus because the methods are *locally* sensitive. For this reason, features such as the unexpected increase and decrease in risk seen at high exposure levels in Figure 4(c) are informative, and provide us with an opportunity to better understand both the strengths and limitations of the data. Such a feature may simply be an artefact of limited data. In general, however, we should strive to keep an open mind because unexpected results may also be clues.

Two major limitations of non-parametric risk regression models should be noted. Both reflect the difficult nature of model selection. The first limitation is computational. The second more fundamental limitation reflects the intrinsic indeterminacy of model selection approaches based on cross-validation.

Our example illustrates that with just one splined variable, the AIC or cross-validation curve $\mathrm{GCV}_{\mathrm{ap}}(\lambda)$ may have more than one local minimum. This observation suggests that it would be difficult to estimate the optimal degrees of freedom for two or more splined variables simultaneously, e.g. to minimize a multivariate cross-validation surface $\mathrm{GCV}_{\mathrm{ap}}(\lambda_1,\ldots,\lambda_p)$, because results of efficient numerical algorithms to find a minimum will depend on the starting value that is used. We can afford to do a brute-force grid-search for a single tuning parameter, but the problem becomes exponentially more difficult as we include more additive components.

Logistic regression models based on the smoothing spline are particularly computer-intensive, even by today's standards. Given a fixed λ, the penalized log-likelihood is maximized using a Fisher scoring algorithm that requires 5–7 cycles for convergence. Within each cycle, parameters for the conventional and splined variables can be estimated using a backfitting algorithm [7] that also requires 5–7 cycles for convergence. Within each of these 25–49 sub-cycles, we are able to estimate the parameters of the smoothing spline in linear $O(n_t)$ time, where n_t is the number of distinct values of the splined variable t, using the ingenious algorithms of Reinsch [19] and Hutchinson and de Hoog [20]. Nonetheless, ideally n_t will be large, namely several score up to a few hundred values, since otherwise this defeats the purpose of trying to estimate $f_{\mathrm{S}_\lambda}(t)$ non-parametrically.

To this point we have determined the fit for a single value of λ. The entire algorithm needs to be repeated about 50 times to chart the cross-validation curve and determine its global minimum. Therefore, this approach requires us to fit about 1250 smoothing splines $(50 \times 5 \times 5)$ during intermediate stages of the procedure. Although each individual smooth can be done efficiently, the total amount of computation taxes the capabilities of microcomputers available in 2002.

The regression splines are less computationally intensive but have other limitations. First, conditional on the number of knots in the model, it is unclear whether the results are sensitive to perturbations of the knot locations. Second, models with relatively few degrees of freedom, such as the best-fit models selected by AIC in our example, may be sensitive to values of a few highly leveraged observations. Finally, even with a moderate number of degrees of freedom, regression splines are less local than smoothing splines with the same df [7, Section 2.8]. However, despite being computationally more tractable, model selection remains an analytical challenge.

Currently, an objective basis for model selection—such as that provided by AIC or GCV—derives from the principle of cross-validation. One constructs an estimate of mean squared error (bias-squared plus variance-squared), and then makes a trade-off between bias and variance by minimizing this estimate. Unfortunately, the need to make a trade-off too often results in a dilemma. Either AIC or GCV may sometimes present two local minima, as in our analyses based on cubic regression and smoothing splines, and sometimes the two values may be very similar, as in our analysis based on linear regression splines. This reflects an intrinsic difficulty inherent in making a bias–variance trade-off: a low-bias high-variance model may have nearly the same AIC or GCV score as a high-bias low-variance model.

How does one deal with this dilemma? A reasonable approach is to present all essentially equivalent models and possibly, other models that provide a sensitivity analysis, as we have illustrated in Figures 4 and 5. This provides an objective framework for negotiating the dilemma, and it is reproducible. However, because there may not be a uniquely best model, it is doubtful that fully automatic methods will make obsolete the need for skilled data analysis and careful scientific interpretation.

We also believe that an alternative approach to model selection—namely not doing some sort of cross-validation and relying solely on intuition—is potentially misleading. In our experience, trying to pick a 'reasonable' or 'plausible' model is difficult, and one may spend considerable effort interpreting 'features' that turn out to be noise. On balance, we believe non-parametric risk regression presents a valuable extension of the standard logistic regression model for case–control studies. However, it is our opinion that the most scientifically credible application of this approach requires a disciplined approach to model selection.

REFERENCES

1. Seigel DG, Greenhouse SW. Multiple relative risk functions in case–control studies. *American Journal of Epidemiology* 1973; **97**:324–331.
2. Greenland S. Dose–response and trend analysis in epidemiology: alternatives to categorical analysis. *Epidemiology* 1995; **6**:356–365.
3. Weinberg CR. How bad is categorization? *Epidemiology* 1995; **6**:345–347.
4. Greenhouse SW. Some epidemiologic issues for the 1980s. *American Journal of Epidemiology* 1980; **112**:269–273.
5. Day GL, Blot WJ, Austin DF, Bernstein L, Greenberg RS, Preston-Martin S, Schoenberg JB, Winn DM, McLaughlin JK, Fraumeni Jr JF. Racial differences in risk of oral and pharyngeal cancer: alcohol, tobacco, and other determinants. *Journal of the National Cancer Institute* 1993; **85**:465–473.
6. Greenhouse SW. Some reflections on the beginnings and development of statistics in 'your father's NIH'. *Statistical Science* 1997; **12**:83–87.
7. Hastie TJ, Tibshirani RJ. *Generalized Additive Models*. Monographs on Statistics and Applied Probability. Chapman & Hall: London; 1990.
8. Hall P, Marron JS. Local minima in cross-validation functions. *Journal of the Royal Statistical Society*, *Series B* 1991; **53**:245–252.
9. Green PJ, Silverman BW. *Non-parametric Regression and Generalized Linear Models*. Monographs on Statistics and Applied Probability, Chapman & Hall: London, 1994.

10. de Boor C. *A Practical Guide to Splines*. Applied Mathematical Sciences. Springer: New York, 1978.
11. Smith PL. Splines as a useful and convenient statistical tool. *The American Statistician* 1979; **33**:57–62.
12. Akaike H. Information theory and an extension of the maximum likelihood principle. In: *Second International Symposium on Information Theory*, Petrov BN, Csàki FC (eds). Adademia kiadó: Budapest, 1973; 267–281.
13. Stone M. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society*, *Series B* 1977; **39**:44–47.
14. Stone M. Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society*, *Series B* 1974; **36**:111–147.
15. Easton D, Peto J. Re: 'Presenting statistical uncertainty in trends and dose–response relations'. *American Journal of Epidemiology* 2000; **152**:393–394.
16. Figueiras A, Cadarso-Suarez C. Application of non-parametric models for calculating odds ratios and their confidence intervals for continuous exposures. *American Journal of Epidemiology* 2001; **154**:264–275.
17. Cleveland WS, Grosse E, Shyu WM. Local regression models. In: *Statistical Models in S*, Chambers JM, Hastie TJ (eds). Wadsworth and Brooks/Cole: Belmont, CA and Pacific Grove, CA, 1992: 309–376.
18. Abrahamowicz M, Du BR, Grover SA. Flexible modeling of the effects of serum cholesterol on coronary heart disease mortality. *American Journal of Epidemiology* 1997; **145**:714–729.
19. Reinsch CH. Smoothing by Spline Functions. *Numerische Mathematik* 1967; **10**:177–183.
20. Hutchinson MF, de Hoog FR. Smoothing noisy data with spline functions. *Numerische Mathematik* 1985; **47**:99–106.